A Critical Analysis of the Fugitive Emissions Management Program Effectiveness Assessment (FEMP-EA) Report

Kyle J. Daun, PhD. P. Eng., FASME
Professor, Department of Mechanical and Mechatronics Engineering
University of Waterloo

Erica Emery, MSc., P. Eng.
Senior Research Engineer, Mining & Energy
Saskatchewan Research Council

David Risk, PhD.
Professor, Department of Earth Sciences
St. Francis Xavier University

Background of FEMP-EA and Purpose of this Report

The FEMP-EA study was designed to study leak and vent sources in a region of Alberta, as well as the durability of leak repairs. It was designed to address several "big gaps" identified in PTAC-led Methane Roadmap Sessions, as described in a presentation to the Environmental Services Association of Alberta in April, 2018:

- Where do leaks occur, what causes leaks, and how do we prevent leaks from occurring (best operations/maintenance practices)?
- Do small leaks turn into large leaks, if undetected or left unrepaired?
- How frequently do LDAR surveys need to be completed, in order to "cost-effectively" manage to the regulatory required levels?
- Does the frequency differ for each facility/source type, vintage, geographical area, etc.?
- How can "super-emitter" leaks be rapidly detected and mitigated?
- What equivalent alternative leak detection technologies and methods exist (or are emerging)?

The focus of the FEMP-EA work was on equipment and sources, with some attention to identifying emerging measurement tools. Fittingly, the authors describe it as a study to "characterize spatial and temporal differences in methane emissions from oil and gas facilities subjected to leak detection and repair (LDAR) surveys through field measurements". The FEMP-EA field study was carried out between August 2018 to May 2019. The study consists of over 4,000 measurements of nearly 200 sites in the Red Deer area. A wide range of facilities were investigated, ranging from very small emission sources from valves, to very large emissions from tank thief hatches, and these results are used by the authors to provide a "snapshot" of methane emissions from upstream oil and gas facilities. Measurements were also carried out over a range of conditions.

These surveys were done almost exclusively using quantitative optical gas imaging, specifically the FLIR GF320-QL320 system. This equipment is quite new, and, while not the focus of this study, a secondary objective appears to be to showcase the capabilities of this technology. Indeed, the authors are using this report to advocate the use of QOGI to quantify large emission sources from tanks (e.g., pp. 18). On the other hand, the novelty of this equipment means that its accuracy and effectiveness is still somewhat undetermined. Moreover, while the manufacturer has provided some "best practices" and there are specific QOGI standards in certain jurisdictions, there are many other scenarios that have not been thoroughly vetted.

While the authors have undertaken considerable effort in carrying out their surveys, a number of concerns have been raised regarding the study and the final report that impair the reliability and usefulness of these results. The objective of this review is to provide objective and constructive criticism of the report, which the authors can use to correct these issues and ensure that Canadian petroleum producers, and the public at large, can derive maximum benefit from the information presented in the report.

To draw robust, defensible conclusions, the data quality must be carefully considered, with a transparent discussion of the measurement uncertainty and outliers, and the report edited such that the conclusions follow clearly from the discussion.

The following sections provide a detailed summary of issues that, in our opinion, need to be addressed in the FEMP-EA report.

1. Typographical and Grammatical Errors, Unit Systems

While the report is generally well-written, it could benefit from a table-of-contents. It also contains many typographical and grammatical errors. These include inconsistent abbreviations, poor punctuation, and additional and missing words. The report should be carefully revised and these errors corrected. For example: on pp. 42, an enumerated list should have numbered paragraphs (1., 2., etc.). Issues with lists and numbering persist through the report. Poor grammar, inconsistent formatting, etc., leads one to the impression that the report has not been carefully reviewed.

The report also uses a number of different units for flow rates and in both SI and US systems (scfm, m^3/day, etc.). It would be good to choose a "common" flow rate unit, and use it to report values parenthetically after other units are used, in order to facilitate comparison of different flow rates.

The authors should carefully review the report and correct typographical, grammatical, and formatting errors. Emission rates should be quantified using a consistent set of units. The authors may also consider adding a table of contents to improve the readability of the report.

2. Uncertainty Quantification

The analysis presented in this report relies exclusively on flux estimates obtained from the FLIR GF320/QL320 system, which, as noted above, is a relatively new system. In their report the authors note numerous times that they recommend further study to better understand the accuracy and precision of QOGI measurements. *This does not absolve the authors from providing uncertainty estimate*, since it is impossible to interpret the results of this study in a meaningful way without understanding the uncertainty attached to flux estimates. This is especially important given the fine-grained nature of the analysis. Accounting for uncertainty will likely change some of the conclusions.

The only attempt to quantify the uncertainty attached to these measurements is a very dubious statement on pp. 51, "These independent tests [AMFC, SRC] have demonstrated the accuracy of aggregate emissions with an average error of about 18%." A careful reading of the AMFC and SRC reports finds a much more circumspect conclusion about accuracy. The "18%" cited in this report seems to correspond to Figure E1 in the AMFC report, which shows that a line fit to QOGI-inferred fluxes from 100 controlled releases of between 20 scfh (9.44 slpm) and 10,000 scfh (4719 slpm) has a slope of 0.82 when plotted against ground truth emission rates, with the average QOGI estimates consistently underestimating the ground truth value. The slope of 0.82 (really between 0.73 and 0.92 with 95% confidence), is obtained by pooling results from the 5 ft and 15 ft release height. Considering only the 5 ft release height, the maximum likelihood estimate for the slope is 0.67, or between 0.54 and 0.79 with 95% confidence. Using the problematic terminology of this report, this translates into an MLE "error" of 33% - much larger than 18%. The SRC report shows an error ranging between -13% and 60% for flow rates between 1-10 lpm (1,000-10,000 cm^3/min), with an average absolute error of approximately 20%.

Moreover (as discussed below) the authors of this study carried out measurements beyond the maximum measurement capability of the camera as specified by the manufacturer, i.e. 360 m^3/day. Were the regression line calculated using only these measurements, the slope would have much lower significance: R^2 value falls from the reported ~0.8 to a much lower value of ~0.2, suggesting that the reliability of the camera may be lower for larger leaks outside of the measurement range. *Accordingly, claiming that the average error of QOGI-derived flux estimates is "about 18%" is a very misleading statement.*

The authors should, at a minimum, acknowledge the various sources of uncertainty in QOGI-derived flux estimates. These include: simplification inherent to the GF320/QL320 measurement model; errors caused by erroneous input parameters (air temperature, wind speed, distance to plume); artifacts induced by reflections, non-uniform background illumination, background motion, natural convection of

water vapor and temperature-induced variations in the refractive index of air; and measurement noise. This list is by no means exhaustive. Some of these error sources are listed in academic publications by the lead author, but they should be mentioned here, and there should be some attempt to identify specific measurement scenarios where these errors may become significant (*cf*. Sec. 3). In extreme cases, some measurements should be removed from the study (*cf*. Sec. 4.1).

The authors cite the AMFC report as a means of verifying the performance of QOGI. However, the AMFC study was conducted in an open and undeveloped site. Does this truly transfer to FEMP-EA where, for example, the indoor spaces or where glint/reflections/accumulating plume videos are seen frequently? Many FEMP-EA measurements are conducted inside of sheds or other buildings. A not-insignificant portion of the videos had glint/reflections and accumulating plumes. Were AMFC measurements done within camera lens focal ranges, and with correct geochemical composition? Are the AMFC error rates or averaging recommendations directly transferrable to FEMP-EA? Maybe this is true for some measurements, but the authors make sweeping generalizations which would not be supported by anyone who understands how these technologies work, or who have read about their susceptibility to error in difficult conditions

The authors hint at additional validation measurements made with a Bacharach Hi-Flow sampler on pp. 52 of their report: "Third, we observe a discrepancy between measurements from QOGI and the Bacharach Hi-Flow sampler in this study. While the accuracy of QOGI estimates in aggregate has been verified through controlled release tests, this discrepancy between QOGI and the hi-flow sampler is an issue worthy of further investigation." From analyzing the data provided with the report, it appears that some Hi-Flow measurements were carried out, and they were sometimes at odds with QOGI-inferred values. However, the report does not present any results from the Bacharach Hi-Flow sampler. Indeed, Section 2.3 (pp. 17-18) seems to imply that this equipment is not used because it is unreliable and cannot be applied to measure emissions from inaccessible locations. This issue should be clarified in the report, and any discrepancies between these two measurement techniques should be laid bare.

The authors should also incorporate conservative estimates of the overall measurement uncertainty, which should be scenario specific. (A global "18%" is not a conservative estimate.) Any conclusions drawn from emissions measurements should be done in the context of these uncertainties; in some cases it may not be possible to draw a robust conclusion, which is much better than making a dubious and unsupported claim.

3. Deviations from Recommended QOGI Measurement Procedure

Although the Alberta Energy Regulator doesn't provide specific instructions related to use expectations for QOGI, the British Columbia Oil and Gas Commission Fugitive Emissions Management Guideline Version 1.0 (July 2019) outlines some requirements for QOGI use in Fugitive Emission Management Programs (slightly paraphrased below to shorten, and with units converted):

- Follow manufacturer written specifications for the specific device
- Use a tripod to steady the camera
- Leak video should be collected for a minimum of 120 seconds for auditability
- QOGI can be used to measure leaks between 0.432 m^3/d and 432 m^3/d

The expectations of a measurement study like FEMP-EA are those of audit-level accuracy and precision, especially since the articulated intent is to disseminate the results in a peer-review journal. The guidelines provided by the BC Oil and Gas Commission are, however, in-line with manufacturer requirements.

The manufacturer's guidelines stipulate the following:

(a) Sufficient Delta Temperature
The infrared radiation contrast between plume and background is an important pre-condition for QOGI measurement. Providence training stipulates a minimum temperature difference of 2 K.

(b) Tripod
Algorithms infer the velocity of the plume from the apparent motion of the plume relative to a stationary background, so a tripod must be used to steady the camera.

(c) Ten replicates
A measurement is to be derived from analysis of 10 independent recordings. As explained by a Providence technical representative: "The purpose of taking replicate readings is to average out the plume behavior. To do that you must capture multiple files of the same leak at different times (with the version you referenced). Our method measures the flux of the plume across a boundary that is a fixed distance from the leak point. Sometime the rate at which the plume crosses our flux boundary is not well correlated to the leak rate (especially when the wind is light and variable). Multiple readings will tend to average the plume behavior." The current BC OGC July 2019 FEMP Guideline specifies 120s of video as the standard for audit-quality measurements.

(d) Within lens focal length-appropriate distance range
OGI cameras can be outfitted with one of three different lenses, each of which is to be used for a certain range of imaging distances between the plume and the camera. The 23 mm lens is to be used for 5 ft to 54 ft (1.5 m to 16.5 m), 38 mm from 8 ft to 90 ft (2.4 m to 27.4 m), and 92mm lens from 20 ft to 210 ft. (6.1 m to 63 m).

(e) Presence of wind and a blowing plume, or an accumulating plume
In the presence of wind, plume velocity is high, which improves quantification accuracy. In windless settings where the plume accumulates, quantification is more difficult. The training video provides an example of a source that was measured once under blowing conditions and again under accumulating conditions, where the accumulating conditions led to a high-biased estimate by about 3x.

(f) Sensitivity
The camera should be operated in normal sensitivity mode whenever possible.

(g) Gas composition
The QOGI measurement is sensitive to gas composition, since hydrocarbon molecules absorb and emit radiation with different efficiency. If the gas does not consist of pure methane, a response factor must be used to account for this effect.

(h) Operating parameters
The emission rate must not exceed 360 m^3/d (250 L/min). The maximum measurement distance is 36.4 m (with 92 mm lens).

Unfortunately, the FEMP-EA draft report offers few methodological details on field measurement practices that can be compared against manufacturer recommendations or Canadian regulatory expectations. On this basis it is difficult to understand field practices and to judge the rigour of the measurements (although some can be discerned by examining raw data in further sections). Accordingly **the authors should clearly define the measurement procedure used to obtain the videos used to infer emissions rates.**

The supplementary data provided with the report reveals that a large proportion of the reported emission rates exceed the operating limits for the instrument. The operating limit is 360 m^3/d, yet values up to 9700 m^3/d (**27 times higher than the operating range**) are reported in the study. The authors seem unconcerned about operating outside of both the manufacturer-recommended range and the range stipulated by the BC OGC guidelines. On pp. 51 the authors state "QOGI improves the range of measurement capabilities, from relatively small emissions (< 10 m^3/d [6.94 slpm] to over 1000 m^3/d [694 slpm]) while the Hi-Flow sampler is limited by the maximum displacement of the blower (650 standard cubic feet per hour [307 slpm])."

What independent evidence, for example multipoint calibrations as shown in Providence/FLIR Advanced Training documents, is available to justify reporting values outside operating range?

The committee examined two measurement subsets in detail: ~150 with above-factory-calibration emission rates, and a further 150 were randomly selected from the remaining data points. A large number of the measurements appear to be carried out in a manner that is inconsistent with recommended practice. The table and notes below provide a general summary of these issues.

Property	**QL320 Best Practice**	**Observations**
Minimum range	The minimum operating range according to British Columbia Oil and Gas Commission (BCOGC) is 0.432 m^3/day**.**	369 raw measurement entries (~10% of total number, or <0.1% of total emissions) were below the factory recommended minimum measurement threshold**.**
Maximum range	The maximum operating range is 360 m^3/d (250 L/min).	139 raw measurement entries (~4% of total number, or ~57% of total emissions) exceeded the factory recommended maximum measurement threshold, for which the tech may still work but the factory

		recommends a user-built calibration and user validation. AMFC may constitute validation for measurements performed in a similar season and under similar outdoor condition, but it is up to the authors to strongly defend (i.e. to provide proof of) extensibility of AMFC results to other seasons or indoor conditions, etc.
Wind	Sufficient windspeeds are required to avoid accumulation of the leak around the boundary.	Many of the leaks accumulated around the boundary, due to being indoors or having low wind. The manufacturer specifies (and provides examples in the training materials) that such conditions may sometimes bias the leak rates upward.
Ambient temperature	All properties of the plume must be accurately imputed into the software.	Ambient temperature seemed suspicious in some cases; for example, one entry recorded -4^{o}C at 10:43 am in August.
Tripod use	Tripods must be used for all measurements.	Tripods were used for most but perhaps not all as some images displayed shakiness.
Delta Temperature	A delta temperature of at least 2-3 ℃ between the plume and the background is required to create enough contrast to quantify the leak.	Sufficient delta temperature was observed in all cases.
Time	The manufacturer and BCOGC recommends QOGI videos be a minimum of 120 seconds, or that 10 short screen captures of ~10 seconds be used for quantification.	Many measurements represent an equivalent 1-minute capture as needed but others are shorter and maybe 20-30 second equivalent (insert in Time part of the table.
Framing	The leak must be positioned at the center of the boundary ring. Most of the plume must be highlighted during quantification and other objects that are not the plume should not be highlighted.	In many cases, the leak was not at the center of the ring or not in the ring at all. For many observations, only parts of the plume were highlighted or other objects (ex: clouds) were highlighted. The camera position was good.

		for the most part, and the center field of view was more commonly affected when multiple different sources are present in the camera frame.
Composition	The camera response function should reflect the mass-emission-rate-percent-weighted response factor (RF). RF should be determined using the response factor calculator.	The plumes were defined as 100% methane although Red Deer is characterized as 82% methane by Johnson et al. 2017. The RF will likely be higher than that of pure methane.
Plume	The plume should be clearly visible. Accumulating plumes are difficult to quantify accurately, according to the manufacturer	The plume was not visible in 13% of the 150 random measurements. The predominant plume transport mechanism for the randomly selected data was 45% blowing, 12% accumulating, 30% mixed, 13% Nan. 54% were point, 35% diffusive, 11% mixed Plume polarity was incoorect in some images.
Backgrounds	Backgrounds should ideally be free of reflections. Outdoors preferred to indoors. Operatory must avoid having other objects in the background, particularly ones that are hot, reflective, or that move.	One-third to one-half of all videos were taken indoors. Background reflectiveness in images was predominantly from H_2O/cloud interference for exterior measurements, and surface reflectiveness of different materials for interior measurements. Some of the interior backgrounds were highly complex and reflective. Background stability was sometimes affected by natural interference such as trees, clouds, and wind moving objects for exterior measurements. For indoor measurements movement was due to moving objects such as labels/tags, and other equipment such as fans in some images.

		The highest 15 measurements were all reported for outdoor settings with hot or very hot background.
Distance and Lens	The manufacturer specifies a usage range for each lens focal length, because at distance there are fewer pixels to utilize when quantifying the gas leak and accuracy suffers. The permissible range of each lens is as follows: • 23 mm lens: 5-54 ft • 38 mm lens: 8-80 ft • 92 mm lens: 20-210 ft	Almost one-third of the measurements reported above the factory calibration range were acquired at a distance larger than is recommended for the 23 mm lens, and in some cases the camera was used at several multiples of the recommended distance. The lens used was frequently not recorded, so we could not always verify.

We also noted database spelling errors and inconsistencies in file notation.

In summary:

- 139 raw measurements (~3.7%) are above the manufacture-specified limit of 360 m^3/day
- 369 raw measurements (~9.8%) are below the BCOGC's minimum of 0.432m^3/day

Measurements that exceed the upper limit are a particular concern. Approximately ~57% of the total emissions reported in the raw files correspond to the 3.7% of measurements that exceed 360 m^3/day. Accordingly, many of the key findings of the FEMP-EA study rely on a small subset of measurements that are outside the established measurement capabilities of the device. Accordingly, it is imperative that each of these measurements be carefully and critically assessed before they are included in the final analysis.

Also notable are some very suspicious ambient temperatures reported in the raw data files. The radiation emitted by the gas depends on the product of the spectral absorption coefficient, which is proportional to gas concentration, and the blackbody intensity, which is a function of temperature. Since these quantities appear as a product, it is impossible for a single-channel infrared camera to simultaneously infer gas concentration *and* temperature. Accordingly, the GF320/QL320 system relies on the operator-specified temperature to connect the measured intensity to the gas concentration, and any errors in this temperature will have a profound effect on QOGI-inferred concentrations and emission rates. Inspection of the raw data reveals a very large range of ambient temperatures, some of which appear highly suspicious; for example, a temperature of over 70°C in December, and a midmorning temperature of -4°C in August.

The impact of excursions from recommended practice on the emissions estimates should be carefully considered and included in the uncertainty estimate. The data should be carefully inspected, and scenarios which are likely to present questionable emissions fluxes should be removed. Ideally, emission rates should also be adjusted to reflect the specific composition of the gases being measured in this study. If this is not possible, the authors should acknowledge gas composition as an additional source of uncertainty.

The authors should also carefully consider the plausibility of QOGI-emissions rates, and flag any suspected outliers. This must be done in a clear and transparent way, using a well-justified set of criteria (cf. Sec. 4.1.).

4. Statistical Analysis

4.1. Data Management and Outlier Removal

The report states that all values were kept unaltered and included in the "aggregate" approach. However, comparison of the raw data files forwarded to PTAC and final project team Excel summary sheets reveal significant discrepancies. These discrepancies raise serious questions as to the degree to which the project team was transparent and systematic in the application of data quality control and filtering methodologies. The following table summarizes the content of the summary spreadsheet provided by the FEMP EA authors.

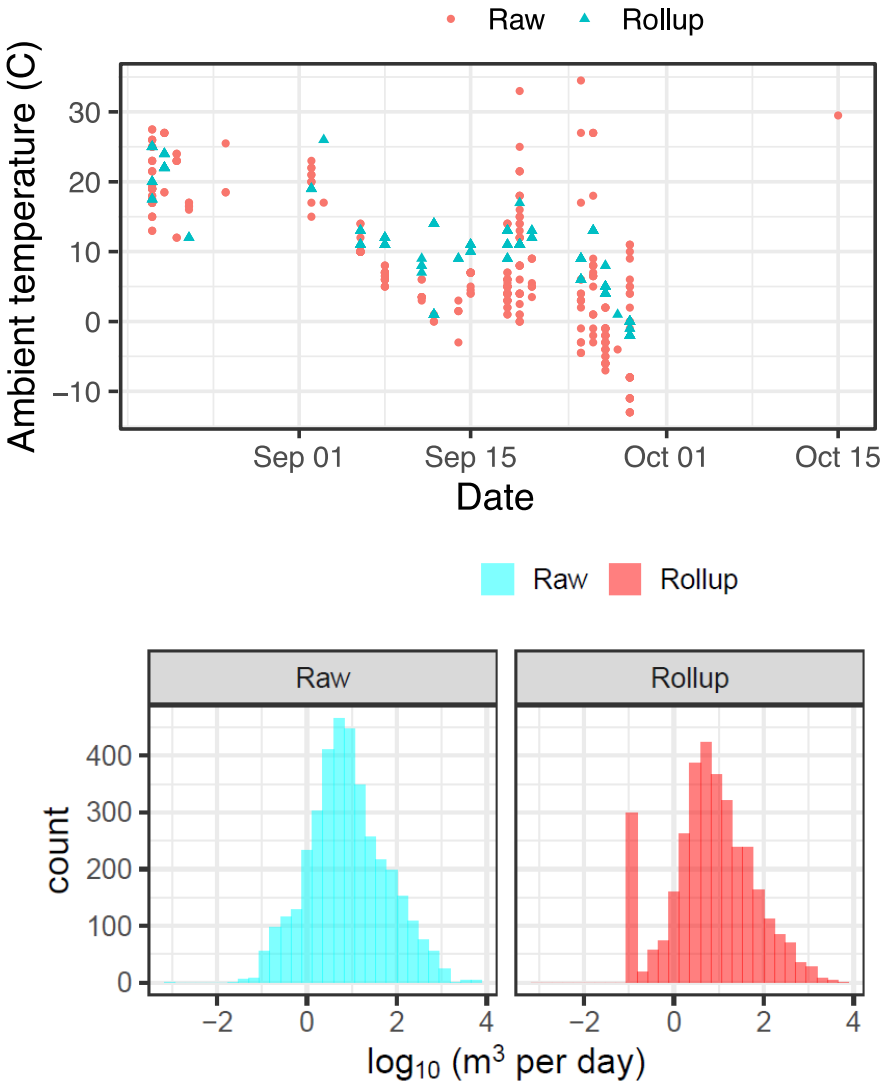
Ranges	August 2018	November 2018	March 2019	May 2019	August 2019
Number of measurements	1117	272	332	480	1317
Leak rate (m^3/day)	0.027-4550	0-308	0-1450	0-44.1	0-4440
Ambient temperature (°C)	-3 – 34	-11 – 12	-5 – 13	10 – 27[3]	-2 – 27[4]
Wind (km/h)	0-41[1]	1-12[2]	1-8	Not recorded	1.2-21.3

[1] 88 entries missing
[2] 212 entries missing
[3] 290 entries missing
[4] 20 entries missing

The raw files contained ~500 additional measurements, and no rationale is provided as to why certain measurements were excluded from the analysis. How was this done? There are some hints in the data that some problematic measurements were removed. For example, the plot below compares the ambient temperatures reported in the raw data files as well as the ones from the summary spreadsheet. The studies that were included in the report correspond to a much narrower distribution of ambient temperatures, which, on the whole, seem to be more plausible than some of the temperatures reported in the raw data files. However, the authors have not provided any insight into the procedure they used to exclude measurements.

A significant number of emission measurements are missing both temperatures and dates. Additionally, there is sometimes more than one entry for one date, which raises questions about how carefully the data was collected and managed.

Comparison of the raw data and summary spreadsheet shows that many of the negative emission rates and higher values have been deleted or nullified. This procedure has a significant effect on the overall distribution of emission rates, as shown below. It is certainly appropriate to remove outliers, including negative emission rates, but these details must be included in the report since it speaks to the limitations of the QOGI procedure.



The FEMP-EA report review committee also discussed whether or not it was appropriate to include emission estimates that were either above (>360 m^3/day, 3.7% of sources) or below (<0.432 m^3/day, ~9.8% of sources) the measurement range of the camera. As noted in Sec. 3, a significant fraction (~57%) of emissions appear to be due to a very small number of sources (~4%), so these sources will have an outsized impact on the overall findings of the report. At least one of these emission rates has been flagged by an industry reviewer as implausible, based on the volume flow rate and the nature of the source.

After careful consideration, the FEMP-EA report review committee recommends that emission rates that exceed the measurement range of the camera should be included in the inventory, provided they are distinguished from the other data, and that the measurements were conducted in a manner consistent with the manufacturer-specified measurement procedure defined in Sec. 3.

The authors should consider supplying a detailed description of some of the largest emitters as an appendix to the report, along with the manner with which the data was analyzed. This information would be helpful when assessing the reliability and relevance of "out-of-range" measurements.

Information could include raw MWIR images, information about ambient temperature and other measurement specifics, and a plot of the individual emission estimates from the multiple visualizations used to construct an single estimate.

*A clear methodology and a set of criteria must be provided that justifies removal of problematic cases and data outliers. It is **crucial** for the integrity of this study that this be done in a transparent way. Specific measurement examples should be specified in the appendices that detail how the calculations are made, particularly for large emitters outside of the manufacturer-specified capabilities of the camera.*

4.2. Error bars

In Sec. 3.4 it is stated that "In all the results presented in the next three sections, the error bars on figures correspond to standard error associated with finite sample sizes and does not include errors associated with individual measurements." What the authors mean to say is that "The error bars represent the variability of the results." This variability may arise from a number of factors that should be explicitly acknowledged. These include: the episodic nature of some leaks; variation in environmental conditions that could influence the size of the leak (e.g., wind speed, ambient temperature, volume of liquid in tanks); and random errors affecting the QOGI measurement. Accordingly, error bars must not be interpreted as an indication in the variability of emission between components and sites, as is presently done in this study. **In doing this, the authors are drawing false conclusions about the significance of their results.**

As an example, pp. 27: "On average, oil sites emit 605 ± 108 m^3/d, while gas sites (excluding gas facilities) emit 404 ± 148 m^3/d." Someone reading this report would reasonably conclude that oil sites actually emit 605 ± 108 m^3/d of methane, i.e., the true emission rate is contained within 497 and 713 m^3/d with, e.g., 95% probability, but this is far from true. This comment applies throughout the report. The authors must refrain from representing their results in this manner.

Moreover, error bars are often calculated using very few samples. As an example, in Figure 10 (pp. 28) categories "gas facility" and "oil MW battery" contain 6 and 9 samples each, so the estimate of the standard deviation obtained with such a low number of samples is questionable. A particularly egregious case is Figure 16 (pp. 31) for "gas facility" and "oil MW battery", where uncertainties are somehow constructed from two estimates. (There are other similar examples throughout the report.) Given that these categories have the largest measured emissions, these uncertainties should be large, and may fundamentally affect the interpretation of these results. In Figure 22, the "gas facility" category has a sample size of *one* and no error bar. The authors need to consider pooling the data into larger categories.

Furthermore, the authors make a compelling case that emission fluxes are not normally-distributed; instead, in almost all scenarios they appear to follow a lognormal distribution. Accordingly, it is inappropriate report the mean and standard deviation of these distributions. Instead, these results should be summarized as box plots.

Error bars should be replaced with box plots that show median and quartile results, and the number of measurements used to construct each result set should be reported along side the data. Results of questionable reliability due to high variability or a very small number of data points should be flagged, and, in some cases removed. Results should not be expressed as X ± Y *in the report unless* Y *is a defensible uncertainty estimate.*

4.3. Statistical Significance of Results

One of the objectives of the FEMP-EA study was to carry out sufficient measurements to derive general conclusions and trends in site-level and component-level emissions. Unfortunately, the authors have taken things too far. In extreme cases inferences are drawn from as few as two samples.

This is especially problematic because it ignores the high degree of uncertainty in the QOGI-derived fluxes (considering both the inherent variability of emission and large uncertainty in individual measurements), and the fact that emissions are dominated by a small number of large emitters. This makes some of the conclusions drawn from small data pools highly questionable. This is acknowledged multiple times in the report, e.g., pp. 18, "It is critical for stakeholders to not directly interpret individual emissions estimates – Monte-Carlo analysis shows that a single estimate can be uncertain by many times the true emissions estimate. Sample size matters – individual quantification estimates can have high uncertainty, while aggregate measurements have low uncertainty. Thus, site-level or operator-level average emissions as estimated by QOGI are more reliable than any individual [estimate]." Unfortunately, despite this warning, the authors do exactly this in their report.

The SRC and AMFC studies showed that QOGI emissions estimates are subject to very large variances, even under very favorable measurement scenarios (good background contrast, good wind conditions, steady leak) are represent a best-case scenario. The variances of estimates from this study are probably much larger. This likely means that, in many cases, the data should be pooled into broader categories, and there should be clear caveats attached to emission quantities reported in the document.

Also, many of these quantities have significant figures that imply an unobtainable level of accuracy. For example, stating that "Fugitive emissions were reduced by 48% at sites where repairs were undertaken within the first four months…" (pp. 2) is obviously not defensible, since the authors cannot report the change in fugitive emissions within a single percentage point. This comment applies throughout the report. (In other words, it could easily be 46%, 50%, or 30% within the expected margins-of-error.) In contrast, statements like "Despite contributing to a significant fraction of the number of emitters (30 – 50%), flanges and valves only contribute at most 15% to total emissions" (pp. 7) appears more reasonable. Again, the manner in which these estimates should be reported will become clear after a more rigorous uncertainty analysis.

Data should be re-interpreted to ensure that conclusions are drawn from statistically-significant results, considering both the number of samples and the uncertainty inherent in QOGI-derived fluxes.

5. Interpretation of the Results

There are a number of places in the report where conclusions do not appear to be supported by the data, or are ambiguous in meaning.

5.1. Fugitives vs. Vents

Although the report distinguishes between fugitives and vents, the respective definitions are buried in the body and not necessarily clear on first read. Terminology for "vents" and "leaks" is defined on pp. 19, but the terms "leaks" and "fugitive emissions" appear to be used interchangeably, and this is not correct. Only some fugitive emissions are leaks. This is a very important distinction, as the implementation of Leak Detection and Repair (LDAR) is designed lower fugitive emissions but will not necessarily have an impact on vented emissions.

Also, by analyzing fugitives and vents in the same graph, the effect of LDAR frequency on leak rates is muddled. Emission results for leaks and vents should be treated separately. Vented emissions are generally larger, and may masking changes in fugitive emissions over time. Analyzing the two types of emissions separately may also allow authors to draw conclusions about the relative contributions of each category.

5.2. Skewness of November 2018 data (Figure E1)

The authors claim that "The skewness of the leak-size distribution reduces with repair, as seen in the November 2018 data where the top 5% of emitters only contributes [sic] to 35% of total emissions." The results from the November 2018 survey are distinct from other surveys, but can not necessarily be attributed to repairs. Notably, while all other surveys show significant emissions from tank level indicators, these are missing in the November 2018 survey. Also, the other surveys were carried out during warmer weather: could this influence either emission rate or the effectiveness of QOGI? Indeed, on pp. 28, the authors state "This reduced skewness could be attributed to several factors including changes in *weather*, *technology performance*, or intervening maintenance and repairs."

5.3. Unclear and Imprecise Reporting of Data Ranges

The authors frequently report ranges of measurements, but it is not clear whether these results are statistically meaningful, and they will almost certainly be misinterpreted by readers. As an example, on Pp. 5: "On average, oil sites emitted 18% to 149% more methane than gas sites…" From Table E3, it appears that they represent the minimum and maximum differences in measured emissions between "gas sites" and "oil sites" over each survey. The average reader would interpret this statement to mean that oil sites emit between 18% and 149% more methane than gas sites with a certain probability (e.g., 95%). This issue applies throughout the report, and must be corrected.

5.4. Pre- and Post-repair Analysis

This section of the report is very important, since one of the main objectives of this study was to assess the effectiveness of LDAR surveys at mitigating leaks, and the frequency with which these surveys should be conducted. Unfortunately, the review committee found this section of the report to be very difficult to interpret.

At the beginning of the section (pp. 45), "Figure 38 shows the observed changes in average site-level leak emissions of sites repaired during each of the 4 follow-up surveys in November 2018, March 2019, May 2019, and August 2019, compared to initial baseline leak emissions measured in August 2018." On pp. 16 the report states that "The three treatment groups simulated typical LDAR surveys at one, two, and three times per year." So, which of these groups are considered in this section? Are they all combined, as suggested in the caption of Figure 38? Based on the methodology section, every site is surveyed in

August 2018, and then some groups are surveyed annually, semi-annually, or three times a year. So, is it possible that some of the sites have been flagged and repaired multiple times, which could bias the results?

Also, the statistical analysis is a real problem here, and this seems to be acknowledged by the authors on pp. 45: "This finding is driven by the small sample sizes of repaired sites in the intermediate surveys which increase the influence of outlier sites in the overall analysis. For example, the March 2019 survey is dominated by one gas MW battery where emissions reduced by over 90% compared to August 2018." First, this finding is significant, yet it is not indicated in Figure 38. Are any of these results statistically-meaningful? It is impossible to conclude whether the improvements are due to repair, environmental factors, or the inherently intermittent nature of some types of emissions. In the case of the gas MW battery in question, what caused the emissions to drop?

The authors then go on to "remove outlier sites" to clarify the overall trends, but the statistical significance of these findings is still not clear without an uncertainty analysis. What criteria was used to remove outlier sites (*cf*. Sec. 4.1)? Most importantly, how do we know that these improvements are due to repairs of leaks? The "improvements" shown in Figure 39 also seem very small compared to some of the error bars reported earlier.

The discussion about "growing leaks" is confusing. The component-level analysis suggests that leaks do not grow when the period between surveys is 6 months, but there is a 30% increase when the duration between inspection grows to one year. The authors then say "…statistically, the increase in emissions occurred in the 6 – 12-month period after the initial survey because components in the 2/year and 3/year groups did not show any increase in emissions." (Given the numerous issues with statistics and uncertainty analysis the authors should be very careful using the term "statistically".) Figure 40 shows no uncertainties, and the use of percent change is misleading. Certainly the change in total emissions over one inspection per year is 31%, while the changes over two per year and three per year are small, but the absolute change in total emissions is not that different between 1/year, 3/year, and total – so, if the measurement error is additive and not proportional, then a good portion of this change could be within the uncertainty of this estimate. Also, the sample size of the 1/year measurements is much smaller than the 3/year measurements, so one would expect a greater degree of variation in the 1/year result. Moreover, even if this result was statistically-significant (which seems questionable), what exactly do they envision happening in the 6-12 month period that doesn't happen in the 0-6 month period?

Section 6.3 "Temporal Analysis" is also confusing. The grey bar indicates the initial average leakage per site based on the initial survey, while the orange bars denote the change in leakage post-repair if the leaks were repaired and the sites were resurveyed 1-4 months, 4-8 months, or 8-12 months after the initial leaks are identified. After an outlier is removed, Figure 41 (b) seems to show a similar improvement in all three scenarios (although no uncertainties are indicated). Does this mean that few new leaks formed over a 12 month duration? We are confused by "…a site that was on an annual schedule (8 – 12 months) might show higher emissions reductions if repairs had been conducted just prior to the second survey and therefore did not have any time for new leaks to appear on site." If the operator is only repairing leaks tagged in the preliminary survey, should it matter when the repairs are carried out? In other words, how would new leaks be repaired if they have not been surveyed? Or, could the operators have used other techniques to detect leaks?

In Section 6.4 "Ideal Scenario Analysis", our understanding from Sec. 3.4.4 (mislabeled 3.5.4) is that the "ideal" scenario of all detected leaks found at any survey are immediately repaired. This is not entirely clear from the discussion. Section 3.4.4 states: "We therefore simulate leak emissions under a simulated 'best-case repair effectiveness' scenario at sites across the 1/year, 2/year, and 3/year survey schedule

by assuming all leaks found in the original baseline survey (August 2018) were repaired. Tags on components help determine whether a leaking component is a new leak or an unrepaired old leak. For leaks that did not have tag, we used the number of leaking components as a proxy. For example, if 3 un-tagged flanges were leaking in August 2018 and 4 were found leaking in August 2019 at the same site, we assume 3 of the 4 were non-repaired leaks from prior surveys and therefore remove them in the ideal scenario analysis."

This description is very difficult to follow. We think that this means that the analysis starts with an inventory of the August 2018 surveys, and identifies the subset of these leaks that were ultimately repaired by the August 2019 survey. It is not clear what happens after that – does this assume that the leaks appear at any time between August 2018 and August 2019, and are identified at some intermediate time depending on survey frequency? This is very confusing, and should be clarified in the report.

In Section 6.5 "Survey Frequency Analysis" the authors state "This 22% reduction could be interpreted as a proxy for the 'natural repair rate' at oil and gas sites – emissions reductions that are likely to be achieved from routine maintenance. However, given the wide variation across operators in site-level emissions, it is likely that this reduction is a result of proactive emissions management from only a few operators in the FEMP-EA study." Again, it is not clear whether this is a defensible statement given the uncertainties involved in the analysis. How many samples are involved in the analysis? Could this drop be due to other factors?

Emissions from vents and leaks should be treated and presented separately. The explanation for the skewness of the November 2018 survey data should be presented in a consistent (and circumspect) way throughout the report. The manner in which ranges of data are reported must be revised to avoid misinterpretation (e.g., do not say oil sites emit between X *and* Y*% more than gas sites unless the interval between* X *and* Y *is statistically-defensible). The section on pre- and post-repair analysis should be extensively revised to explicate and clarify the methodology. Results and inferences drawn from quantitative analysis must consider the uncertainty and statistical significance of the measurements.*

6. Report Conclusions

In the Summary and Conclusions section, the authors make some questionable claims about the impact of their survey:

- The authors claim that their survey "Characterized all emissions sources as fugitive emissions or vents and quantified over 90% of all emissions." How can they justify this statement? One could interpret this statement as "90% of all emission sources were identified from the chosen sites" or "90% of all emissions from these sites were quantified". Neither of these interpretations is likely to be true.
- The authors also state that they "Developed a comprehensive data collection strategy that goes beyond current industry and regulatory standards, as well as the requirements of the study design." It is not clear what this means. How does the data collection strategy exceed the requirements of the study design? Are these study design requirements specified in the RFP or contract? If so, they are not defined in the report. This statement should likely be removed.
- The authors claim that they have "Established the impact of repair process on fugitive emissions reductions across different LDAR survey frequencies." It is not clear that this is actually true, given the significant issues in the statistical interpretation of the data.
- The authors claim that they "Collected data with significant detail and robustness to serve as a template for future ground-based methane measurement studies. We also identify the importance of characterizing and quantifying all sources to appropriate [sic] sample the small number of high-emitting components and sites that contribute disproportionately to overall emissions." Few of the emission rates presented in this study can be characterized as statistically-robust. There is little indication as to the variance in the emissions estimates, and, by drawing conclusions based on highly uncertain QOGI estimates (e.g., small sample sizes) the authors are setting a bad example. If the authors wish to highlight the importance of sample size, given the high variability in QOGI estimates and the fact that few emitters contribute most of the emissions, they should carry out a statistically-rigorous uncertainty analysis. In some cases, a reasonable outcome may be that the result has too much variance to draw any conclusion – this is far more valuable than a suspect emissions estimate.

The above conclusions need to be supported, modified, or expunged from the report.

7. Miscellaneous Comments

The following issues should also be addressed or corrected in the revision:

- On pp. 4 it is stated "the federal and provincial governments have developed regulations to reduce methane emissions by 40 – 45% below 2012 levels by 2025." On pp. 12 it is stated that this reduction must occur by 2035. Which is correct?
- Pp. 17:"…the FLIR camera is qualitatively verified every day before starting the survey using a propane standard at a flow rate of 50–60 gh^{-1} from a ¼ inch orifice…" Is this measurement done in quiescent conditions, or exposed to the wind? In general, the authors should provide more information about the validation methodology, either within the report or as an appendix.
- Pp. 17: "Hourly changes in weather are not as important if the general outlook for the day (sunny, partially cloudy, etc.) remains consistent." It would be good if the report were to elaborate on this point, since this suggests that the QOGI-inferred fluxes may be biased by the weather.
- Pp. 24: "Thus, 55 out of 172 sites surveyed (32%) did not have any leaks". The authors mean to say that these sites did not have any leaks *that could be detected by the QOGI methodology*.
- Figure 7 (and others): Do the numbers above the bars indicate the number of components sampled, or the numbers found to be leaking? It would be good to clarify this. Again, it would be more useful to see a box/whisker plot instead of error bars.
- Pp. 26: "In line with many recent studies, site-level emissions are less skewed than component-level emissions, with the top 5% of sites contributing to 43% of total emissions." Is this a function of the smaller sample size for component level emissions? If so, this should be explained.
- Pp. 28: "Of the 38 sites surveyed, 3 sites did not have any emissions and a further 8 sites did not have any leaks". Again, these sites did not have *detectable* emissions and *detectable* leaks. Are the three sites that had no detectable emissions a subset of the 8 sites with no leak? Or are there 11 with no leaks? This seems to be the case, based on what is written on pp. 38: "Of the 179 sites surveyed, 14 did not have any emissions and a further 31 sites did not have any leaks. Thus, 45 out of 179 sites surveyed (25%) did not have any leaks."
- Pp. 33: "The top 5% of sites contribute to 28% of total emissions, less skewed compared to component-level emissions but in line with scientific evidence from other studies across the US and Canada." This statement should be supported with relevant references.
- Pp. 42: "The various colored stacks represent the relative proportion of emissions associated with each site-type, not absolute emissions. Emissions from gas facilities (dark blue) account for 34% of operator 4's total emissions – the figure should not be interpreted to mean that the average site-level emission for operator 4's large gas facilities is 400 m^3/d/site." The review committee were confused by this explanation. We think the authors are saying that the amount of emissions per site type may be highly skewed. A concrete example would be helpful to clarify this point. (Or, it may make sense to report average emissions per operator, and not per operator per site.)
- Pp. 42: "Three, *average* site-level emissions also exhibit skewed behavior, similar to component-level and site-level emissions. The four operators with highest-average site level emissions (top 20%) contribute to 84% of total emissions – together, they operated about 50% of all sites measured in the August 2018 survey. This finding could pave the way for a more differentiated form of methane mitigation policy, one that depends on the performance of each operator." Is it appropriate for the authors to speculate on regulatory policy, given the mandate of the report? It seems to be outside

the scope of the report. Moreover, it is not clear that the QOGI-derived emission estimates are sufficiently robust to draw conclusions like this.

- Pp. 51: “QOGI can be used estimate all emissions at sites”. This is far from true. This statement must be removed.